Bielefeld Academic Search Engine

Technik und Praxis

Vortrag auf der 8. Inetbib-Tagung in Bonn am 4.11. 2004

Friedrich Summann
Universitätsbibliothek Bielefeld

Im Rahmen dieses Vortrages sollen Informationen zum technischen Hintergrund der BASE-Lösung gegeben werden. Die FAST-Suchengine hat eine streng modulare Struktur. Es gibt vollkommen getrennte Frontend- und Backend-Server. Zur Zeit wird für BASE ein Frontend-Server eingesetzt, wobei leicht weitere Server ergänzt werden können. Ebenso wird zur Zeit ein Backend-Server genutzt, der zu einem Multi-node-System ausgebaut werden kann. Damit ist mit dem FAST-System eine Server-Farm mit Tausenden von Rechnern möglich, wie sie von den gängigen Suchengines bekannt ist. In der Vergangenheit ist diese Struktur bei der FAST-Suchengine Alltheweb auch umgesetzt worden,

Wie die Module miteinander kommunizieren, zeigt die folgende Abbildung:

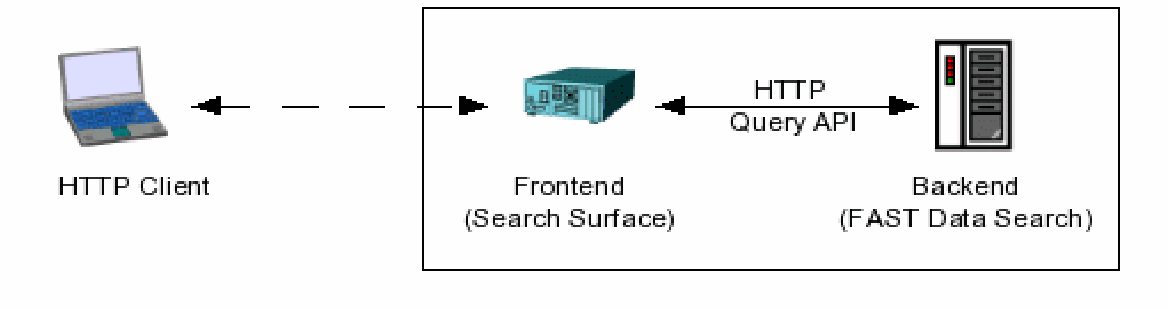

Abbildung 1

Die Aufgaben des Softwaresystems verteilen sich wie folgt auf die beiden Systembestandteile:

Frontend:

- Suchoberfläche (Einfach, erweitert) (PHP)
- Ergebnisermittlung und Ergebnisanzeige

Backend:

- Harvesting (Perl OAI harvester)
- Preprocessing and Datentransfer von BRS,
  OAI-DC und anderen Datenbankformaten mit Perl
- Filetraverser, Crawler
- Dokument-Processing und Indexierung
- Retrieval und Ergebnisaufbereitung

Das PHP-Frontend läuft auf dem Standard-Web-Server der UB Bielefeld, der insbesondere die statischen Web-Seiten der UB bereitstellt. Im Einzelnen gelten die folgenden technischen Details für die Maschine:

- Siemens Primergy, 2 x 800MHz CPU, 1.28 GB RAM
- RAID1, Adaptec SCSI, 36GB
- SuSE Linux 9.0, Kernel 2.4.21-smp
- Apache Web Server mit PHP 4

Da das PHP-Skript keine grosse Maschinenleistung benötigt, ist die Installation auf dieser Maschine unproblematisch.

Für das Backend werden zwei Maschinen verwendet, da ein Testsystem die Neuentwicklung ohne Auswirkung auf das Live-Systems ermöglicht. Seit Ende Oktober 2004 laufen beide Rechner mit FAST Search 4.0.

Für das Live-System gelten die folgenden Kennwerte:

- Intel PC Dual Pentium, 2 x 2.8 GHz CPU, 2GB RAM
- RAID5 + Hotspare 290 GB
- Suse Linux 9.0
- System Report: 344.8GB total, 24.8GB belegt, 320GB free
- 35 Collections, > 700000 Dokumente
- FAST Search 4.0 (PHP 4, Python 2.2)

Im Backend-Bereich werden die Daten erfasst, in das interne Format der Suchengine transferiert und dann indexiert. Dabei kann an verschiedenen Stellen in den Datenfluss (s. Abb. 2) eingegriffen werden. Beim üblichen Datenerfassen von Suchmaschinen, dem Crawlen von Web-Seiten per Robot, ist die Notwendigkeit, Einfluss auf die Verarbeitung der Daten zu nehmen, eher gering. Hier sind ein paar Korrekturen und Ergänzungen bei der Behandlung von Spezialformaten vorgenommen worden. Viel zu tun gibt es allerdings bei der Anpassung der Schnittstelle File Traverser, die im BASE-Rahmen für die Übernahme von Datenbankinhalten und insbesondere bei der Bearbeitung von OAI-Daten verwendet wird. Hier wurden zahlreiche Skripte (Perl, XSLT) geschrieben, die aus den verschiedenen Datenformaten das interne XML-Format erzeugen.

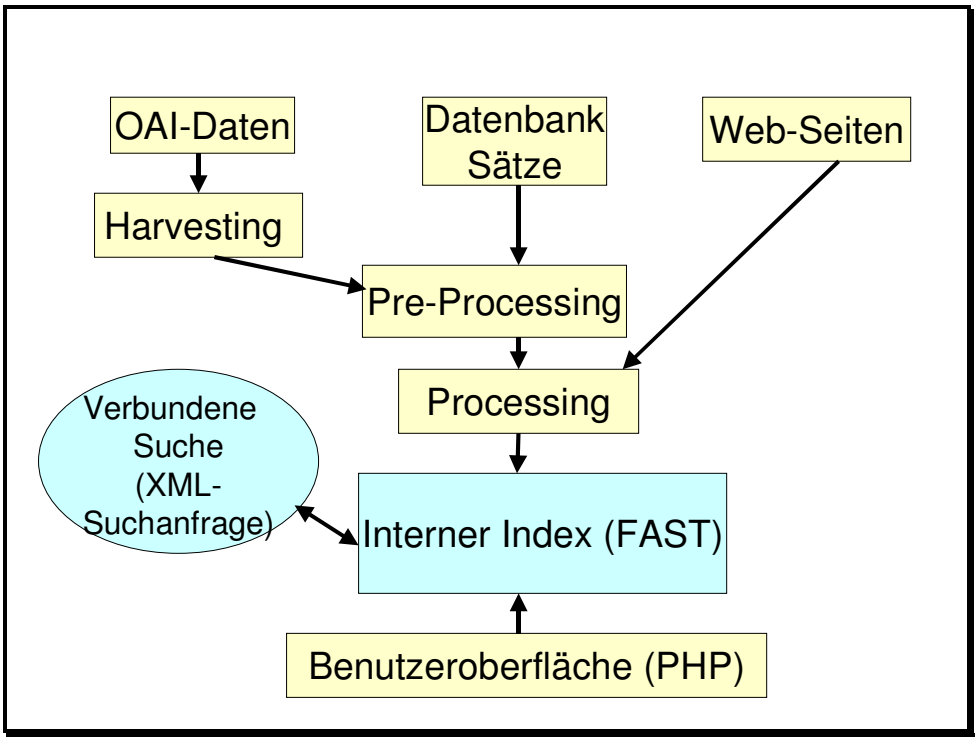

Abbildung 2: Datenfluss bei BASE

Entwicklungsumgebung und Tools

Die für den Erfassungsprozess genutzten FAST-Module und die zusätzlich entwickelten Module und Tools listet die folgende Aufstellung auf. Diese Liste zeigt zudem auf, dass die Eigenentwicklungen sämtlich auf Basis von Open-Source-Lösungen entstanden sind.

- **OAI Harvester tool (für das Harvesten der Metadaten) (open source)**
- **Perl, XSLT (Preprocessing)**
- **Python (Processing)**
- **PHP (Frontend-Programmierung)**
- **Apache**
- **Linux**

Die folgende Tabelle stellt die verwendeten FAST-Tools und die selbst entwickelten Tools zusammen und ordnet sie den einzelnen Arbeitsschritten zu.

	FAST	Zusatz-Entwicklungen
Data Loading	Crawler, File Traverser, DB Connector	OAI Harvester Datenbankexport
Pre-Processing		Perl, XSLT Transferskripte
Processing	Standard stages	Python stages
Indexierung	Indexer	
Retrieval, Navigation	Search API	PHP-Skripte

Indexstruktur

Die Indexstruktur lehnt sich am Dublin-Core-Format an, da die wichtigsten strukturierten Daten im OAI-Bereich in diesem Format vorliegen. Dazu sind ein paar zusätzliche Felder definiert worden, um für Suche und Anzeige wesentliche Informationen intern ablegen und darauf zurückgreifen zu können. Die Erweiterungen betreffen die Felder ISBN, ISSN, die DOI für Resolving-Dienste, das Erscheinungsjahr als Zahl und ein Feld, das festhält, ob Metadaten, Volltext oder beide Merkmale gemeinsam für den betreffenden Satz enthalten sind.

Die Benutzeroberfläche

Das BASE-Livesystem (http://base.ub.uni-bielefeld.de) ist im Juni 2004 öffentlich bereitgestellt worden. Die dabei angebotene Benutzeroberfläche ist durchgängig zweisprachig in Deutsch und Englisch realisiert worden. Neben einer einfachen Suchmaske im Google-Design (Abb. 3) wird eine erweiterte Suche mit ergänzenden Suchfunktionen angeboten.

Abbildung 3: Einfache Suchmaske

Dabei (Abb. 4) werden die ergänzenden Funktionen wie differenzierte Suchaspekte, Kollektionen-Auswahl und Suchhistorie angeboten. Bei beiden Suchmasken lässt sich die Suche auf freie Dokumente eingrenzen.

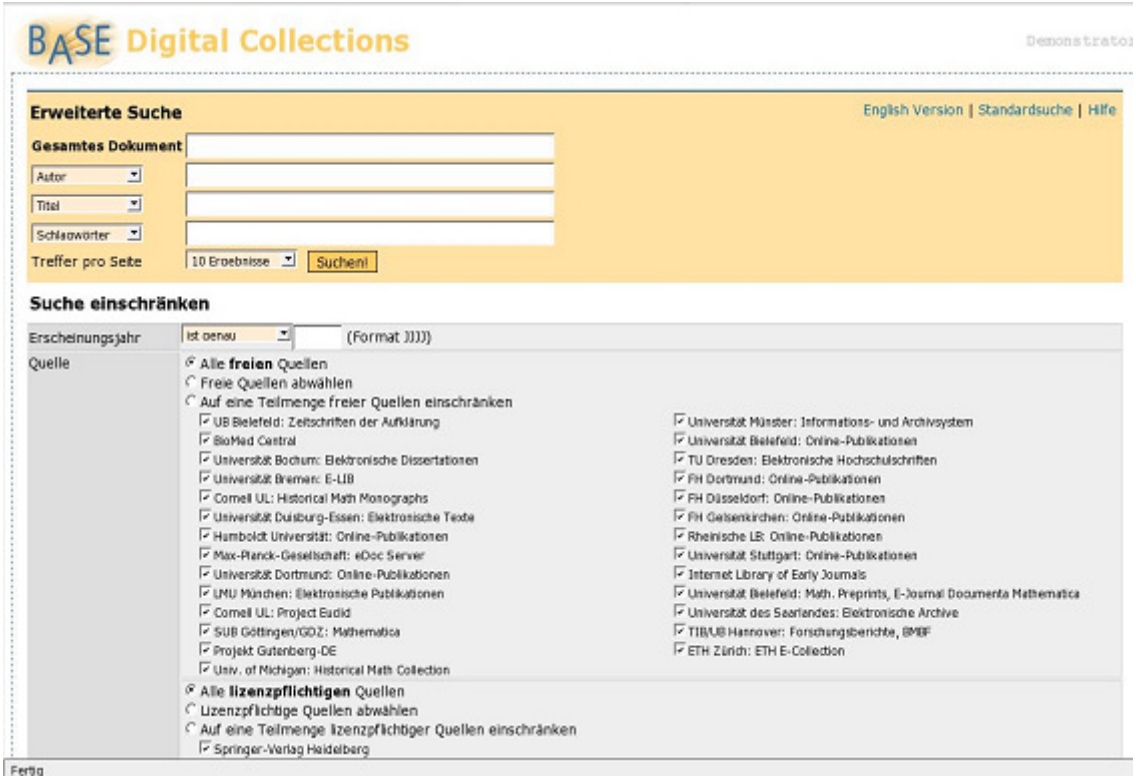

Abbildung 4 Erweiterte BASE-Suchmaske

Die Ergebnisanzeige (s. Abb. 5) unterscheidet sich vom Suchmaschinenstandard durch eine differenzierte Anzeige von Metadaten, wenn solche im jeweiligen Dokument vorhanden sind. Daneben werden Möglichkeiten zur Suchverfeinerung, auf Metadatenebene nach Autoren und Klassifikation und nach formalen Aspekten wie Dokumentformat und Kollektion angeboten. Dabei werden aus der Ergebnismenge die Inhalte aus den betreffenden Feldern extrahiert zu einem Auswahlmenü zusammengestellt. Eine Erweiterung der Suche in Bezug auf das einzelne Dokument ermöglicht es, nach ähnlichen Dokumenten im Gesamtindex (Find Similar), in der Treffermenge (Refine Similar) zu suchen oder gerade die Ähnlichen in der Treffermenge auszuschliessen (Exclude Similar). Die Suchhistorie rundet die Möglichkeiten der Benutzeroberfläche ab.

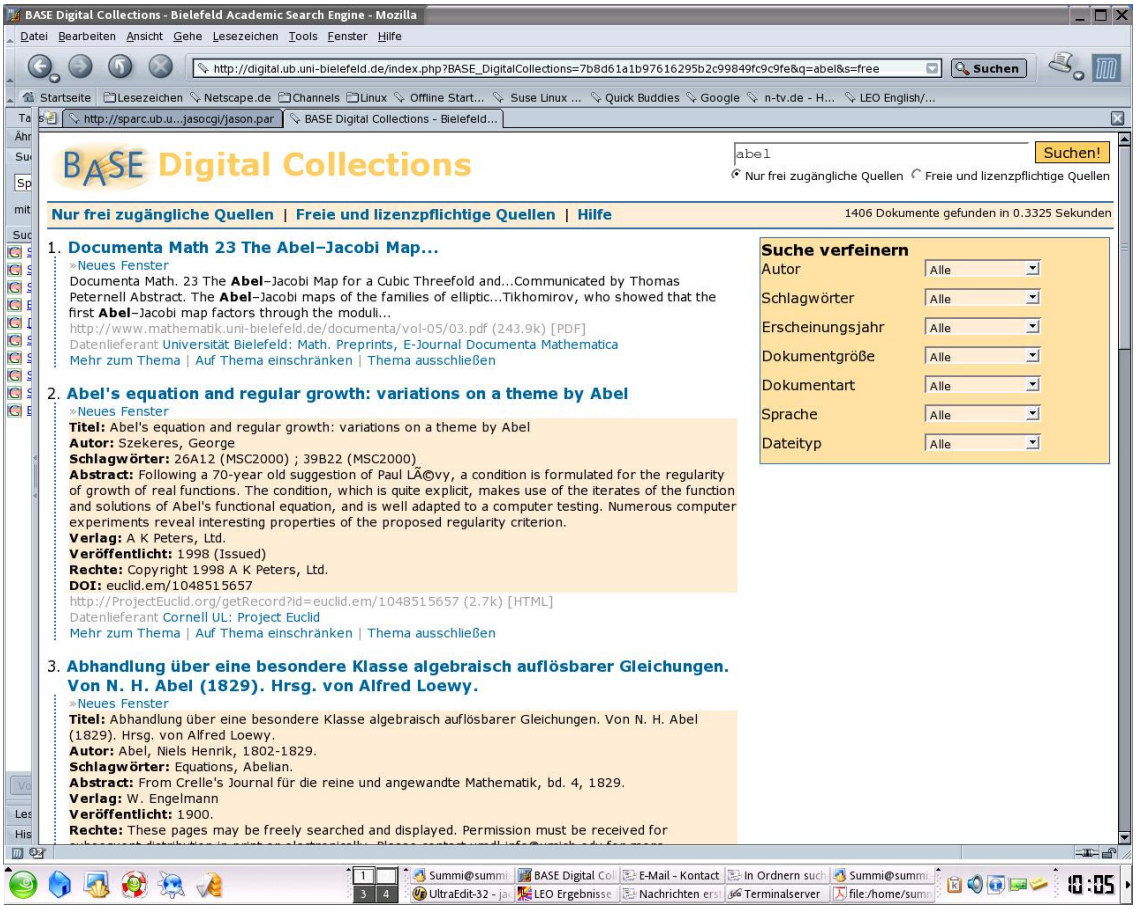

Abbildung 5 Ergebnisanzeige

Für die nächsten Schritte der Weiterentwicklung ist eine Prioritätenliste zusammengestellt worden. Beim Frontend soll mit der Einführung von Templating-Technik die flexible Integration der Suchmaschine in externen Umgebungen unterstützt werden, wobei damit lokale Views auf die BASE-Suchmaschine mit Festlegung von Such- und Ergebnisparametern ermöglicht werden sollen. Schon jetzt ist es möglich, in beliebige Web-Seiten eine oder mehrere Suchzeilen einzufügen, mit der die BASE-Suchmaschine in eine externe Portalumgebung eingebunden werden kann (s. Abb. 4 mit einem Beispiel für die University of Michigan Libraries). Eine Weiterentwicklung in diesem Bereich soll dann auch eine Differenzierung und lokale Anpassung der Such- und Ergebnisanzeigeoberfläche unterstützen. Weiter sollen das Suchinterface auf Basis der Such-API erweitert werden. Dazu gehört auch die Möglichkeit, den Suchindex bestimmter Felder (z.B. Autor, Schlagwort) aufzublättern. Zudem soll bei der Ergebnisanzeige eine Kombination aus Metadatenanzeige und zugehörigem Volltext realisiert werden.

Abbildung 6 Einbettung in externe Suchumgebung

Im Bereich des Backend wird ein wichtiger Schwerpunkt auf der Automatisierung und Konfigurierbarkeit der Harvesting- und Preprocessing-Abläufe der Dokumente liegen, um insbesondere die Probleme mit unterschiedlichen Feldinhalten im Bereich OAI-Harvesting in den Griff zu bekommen. Hier sind schon beachtliche Fortschritte erzielt worden, so dass sich die Zahl der eingebundenen OAI-Server in kürzer werdenden Abständen erhöhen liess. Der Bereich Verbesserung der Suchresultate (auf Basis der FAST-Features Ranking, Boosting, linguistische Methoden) soll einer Optimierung nach wissenschaftlichen Kriterien unterzogen werden. Die Verbesserung der Performanz ist immer ein gewichtiges Thema und muss auch unter dem Aspekt zunehmender Nutzung bearbeitet werden. Da die Such-Engine auch Basis-Dienste für externe Systeme und Portale bereitstellen soll, ist die Implementierung von Standard-Schnittstellen (Z39.50, OAI, SOAP) vorgesehen. Im Punkt Zusammenarbeit mit anderen Systemen ist darüber hinaus die Aktivierung der Features Verteilte Suche und Verbindung mit externen Indexen geplant.

Neben diesen Weiterentwicklungen der UB Bielefeld gibt es bei FAST Entwicklungen, die für den Kernel angekündigt worden sind und damit in kommenden Softwareversionen aktiviert werden können. Hier sind besonders die Punkte Ankeranalyse, Analyse der Linktopologie, Zitatanalyse, automatische linguistische Analyse, Pushdienste, Personalisiertes Ranking und sprachübergreifendes Information Retrieval relevant. FAST arbeitet z.B. bei der Entwicklung des praxistauglichen Einsatzes linguistischer Methoden mit dem CIS (Centrum für Informations- und Sprachverarbeitung) der Ludwig-Maximilians-Universität München zusammen.